A SYNOPSIS
on

PREDICTION OF DIABETICS SYMPTOMES USING WEARABLE DEVICES AND AI

To be submitted to the

Mahatma Gandhi Mission University,
Chhatrapati Sambhajinagar, MH, India

For the Award of the Degree

DOCTOR OF PHILOSOPHY

in
Computer Science and Engineering
(Faculty of Engineering and Technology)

By

Madhuri Kawarkhe

Under the Guidance of

Dr. Parminder Kaur

Department of Computer Science and Engineering
Jawaharlal Nehru Engineering College of Engineering
MGM University, Chhatrapati Sambjajinagar-431003,
Maharashtra (INDIA)

July 2025

DECLARATION

I hereby declare that, the present work completed in the form of a thesis entitled, “Prediction Of Diabetics Symptoms Using Wearable Devices And AI” is an original work carried out by me under the guidance of Prof. Dr. Parminder Kaur, and has not been submitted or published in any form for the fulfilment of any other degree or any other similar to this in other University.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis.

Madhuri Kawarkhe

PH21CSE15

Place: Chatrapati Sambhajinagar

Date:

MAHATMA GANDHI MISSION UNIVERSITY

CHATRAPATI SAMBHAJINAGAR

CERTIFICATE

This is to certify that Amita Jajoo has carried out the work embodied in this thesis entitled “Prediction Of Diabetics Symptoms Using Wearable Devices And AI”. The work included in this thesis is original, unless stated otherwise and has not been submitted for other degree of MGM University or any other University. All references to the work of others have been duly cited in the text.

Prof. Dr. Parminder Kaur

TABLE OF CONTENTS

1	**Introduction**	**1**
2	**Summary of Literature Survey and Research Gap**	**3**
3	**Research Objectives and Scope**	**5**
4	**Contributions**	**7**
5	**Proposed Methodology and Result**	**8**
6	**Conclusion**	**15**
7	**Organisation of the Thesis**	**16**
8	**References**	**19**
9	**List of Publications and Conference Attended**	**23**

List of figures

Sr. No.	Title of Figure	Page No.
1	Comparative analysis of ML models on Pima Dataset	11
2	Proposed System Architecture for Diabetes Prediction	12
3	Performance of Ensemble Classifiers on Pima Dataset	12
4	Blood Glucose Prediction Architecture	15
5	Augmentation Results for D1namo Dataset	16
6	Hybrid Models comparison for prediction of blood glucose	17
7	BGL Forecasting of CNN- LSTM model of last 90 Minutes	17

INTRODUCTION

Diabetes is a long-term condition where the body produces minimal or no insulin or is unable to use its insulin well, resulting in high blood glucose (BG). Diabetes is typically diagnosed in a clinic with a series of tests. Results are then used by practitioners to decide on treatment.Despite this, forecasting diabetes accurately is challenging. Current prediction models are incapable of providing correct results due to two primary reasons: lack of enough data and outliers.

Our study consists of two approaches namely Diabetes Prediction and Forecasting of blood Glucose levels (BGL) for Diabetics. Our First approach presents a proposed prediction model using the Pima Indian Diabetes Datasset.

The main aim of this study is to improve the accuracy of diabetes mellitus prediction by utilizing various machine learning techniques, including ensemble methods such as Stacking, Hard Voting, and Soft Voting, with base classifiers like AdaBoost, Logistic Regression, Random Forest, Gradient Boost, Linear Discriminant Analysis, Extra Trees, and Cat Boost. For this experimentation, we will be using the Pima Indians Diabetes dataset, which gathers details on patients with and without diabetes, to construct and evaluate each model before selecting the optimal ensemble model to address this issue. The best performing model was the ensemble model using soft voting. However, the model had a high bias and low variance, which was addressed by calibration.Our study aims to predict diabetes with the highest possible accuracy using the provided data. We've enhanced our model's performance to 90.62% accuracy by employing preprocessing techniques and ensemble methods. To evaluate our model, we used the Area Under the Curve (AUC). This metric is ideal because it effectively measures the classifier's ability to distinguish between different categories and isn't affected by class distribution.This research contributes to the field by demonstrating how to design effective ensemble models for diabetes prediction. We achieved our results by experimenting with three distinct ensemble models.

The prevalence of diabetes is increasing on a global scale every single day. It has become imperative to monitor and manage diabetes mellitus since it can lead to hypoglycemia, hyperglycemia, and long-term consequences like neuropathy, retinopathy, nephropathy, cardiovascular disease. Diabetes care is often generalized, neglecting physiological responses, and lifestyle differences which creates a need for individualized approaches. Individuals living with diabetes have greatly benefited from the recent technological innovations which have upgraded traditional diabetes care from reactive to proactive, helping individuals avoid complications. The evolution of continuous glucose monitoring (CGM) and wearable sensors has generated abundant, rich, and high-frequency data which can be used with ML and artificial intelligence (AI) for more accurate and automated diabetic management solutions.

We are using the D1NAMO data set downloaded from (https://zenodo.org/records/1421616, Published October 19, 2018). We preprocessed the gathered data in a pipeline prior to presenting it to the machine learning (ML) classifiers. We ran extensive experiments for assessing the performance of different preprocessing and feature selection methods on ML as well as deep learning (DL) classifiers.

This project aims to improve diabetic care by leveraging advanced data science and ML techniques to analyze, predict, and manage glucose levels and potential complications.

SUMMARY OF LITERATURE SURVEY AND RESEARCH GAP

AI in the form of ML and DL is revolutionizing diabetes care by developing predictive models that transcend mere diagnosis to predict glucose levels, disease activity, and complications. The models are based on multi-sources of data ranging from routine questionnaires to CGM systems and frequently combined in hybrid or ensemble systems for better performance.

2.1 Diabetes and Glucose Prediction

DL models such as neural networks have shown extremely good accuracy in predicting diabetes. Huma Naz et al. [4, 16] obtained 98.07% accuracy using DL with RapidMiner. Neural network models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs) show good performance. Swapna G. et al. [5] obtained 95.1% accuracy with a CNN-LSTM model on Heart Rate Variability (HRV) signals. For glucose prediction, LSTMs outperformed conventional methods [19], and Bhimireddy et al. [20] discovered that a sequence-to-sequence Bidirectional LSTM (BiLSTM) was the most effective model for CGM data. BG-Predict, a deep multi-component network [22], and a straightforward Artificial Neural Network (ANN) [24] also gave clinically meaningful outcomes for BG concentration estimation.

2.2 Forecasting Complications and Disease Progress

In addition to diabetes risk prediction, scientists are also developing future health outcome prediction models. Zhou and Myrzashova [6] have developed a model predicting not just risk for diabetes, but also future potential subtypes of the disease. For complications, 97.8% accurate models for the prediction of eight common complications were established by Yazan Jian et al. [25]. Cut Fiarni et al. [29] worked with three independent complications (retinopathy, nephropathy, and neuropathy), and they found strong risk factors such as gender, BMI, and duration of diabetes. Mathieu Ravaut et al. [26] have worked on a three-year predictive model for poor outcome from complications, and it is a planning tool for health care resources.

2.3 Hybrid and Integrated System

The usage of more than one algorithm and technology generally improves performance. Han Wu [15] created a hybrid model combining logistic regression and K-means clustering, which was better performing. Jobeda Jamal Khanam et al. [16] concluded that the hybrid model of merging logistic regression and Support Vector Machine (SVM) worked marvelously. For personalized care, M. V. Kabyshev et al. [27] created a client-server framework from Apple's iOS operating system (HealthKit, ResearchKit, CareKit) that relies on cluster analysis and Random Forest models for far-from-patient risk estimation. Likewise, Kim et al. [28] built an individualized DL model based on a simple Gated Recurrent Unit (GRU) architecture for short-term glucose forecasting in hospitalized subjects.

2.4. Research Gap

1. Almost all researchers have used only static PIMA dataset for prediction [14-18].
2. The prediction system must be capable of predicting the possibility of diabetes along with the future level of BG.
3. Decision-making systems have to have sufficiently high performance and be resistant to loads from a large number of different clients [4].
4. There are few studies on the Type-1 glucose prediction model while many studies performed on Type-2 glucose prediction [5].
5. Execution time for model building and validation with feature selection will need to be considered while finding the accurate ML methods [27].
6. To predict the patient’s BGL more focus should be required by considering patients' behavioral, lifestyle, activity and health related aspects [29].
7. Most of the researchers did not focus on the processing of the users’ meal preferences, diabetic education technology, medicine reminder notification [30].
8. Hybrid algorithms are indeed being developed and researched for more accurate and reliable BGL prediction, particularly for individuals with diabetes [31].

RESEARCH OBJECTIVES AND SCOPE

3.1 Objectives of Research:

Diabetes management is a complex, highly individualized process consisting of ongoing monitoring and immediate decision-making about diet, exercise, and medication. Old methods based on past information and sporadic clinical visits are not well suited to deliver anticipatory and pre-emptive care. But with the development of multi-sensor devices like Continuous Glucose Monitors (CGMs) and wearables, patient-specific information has proliferated. This offers a historical opportunity to use machine learning and deep learning to transition from reactive care to a new era of predictive, personalized diabetes management.

The goal of this project is to use these rivers of data to create an integrated system for next-generation diabetes prediction and management. To achieve this, the project has the following master goals:

1. To develop a comprehensive data pipeline for multi-sensor diabetic patient data.
2. To build, test, and compare various ensemble ML models for glucose prediction.
3. To implement and assess various DL architectures for glucose forecasting.
4. To Facilitate patient-specific glucose data analysis and visualization.
5. Augment existing diabetic patient datasets to enhance model generalizability and robustness.
6. To identify key features driving glucose fluctuations.
7. To develop a model to predict the likelihood of complications related to diabetes based on patient data.

3.2 Scope:

study scope is to promote diabetes control and risk estimation through an extensive, data-based process. The research would not only extend as far as a basic diagnosis model and would encompass the complete data capture to clinical risk assessment cycle.

1. Data Engineering and Preprocessing: The study seeks to create an effective data pipeline to gather, cleanse, combine, and transform multi-sensor diabetic patient data. This involves looking into new data augmentation methods for mitigating data sparsity and enhancing model generalizability.
2. Comparative Model Development for Glucose Forecasting: The study will comparatively develop, verify, and contrast conventional ensemble machine learning models and new state-of-the-art deep learning models to determine how they fare in short-term glucose forecasting.
3. Feature Importance Analysis: The study will analyze and measure the most impactful features of the multi-sensor data that have the most contribution towards glucose fluctuations, providing clinical insight into factors that lead to blood sugar variation.
4. Patient-Centered Analysis of Data: The application entails the formulation of glucose data analysis and visualization methods for specific patients to facilitate individualized insight and aid personalized clinical decision-making.
5. Long-Term Complication Prediction: The research will formulate a state-of-the-art predictive model for exploring and predicting the probability of a patient's development of diabetes-related complications and thus shifting the treatment agenda from reactive to proactive and preventive.
6. Insulin Dosage Recommendation System: The project will create a model that uses predicted glucose values and other key patient data (e.g., insulin on board, carbohydrate absorption) to make individualized, timely insulin dose recommendations. This will be a major step toward a clinical decision-support system enabling proactive management of blood sugar and avoidance of both hyper- and hypoglycemic episodes.

CONTRIBUTIONS

Designing a model for predicting BGL offers significant contributions to society, doctors, and diabetic patients. These models, often leveraging ML and CGM data, empower individuals and healthcare providers with proactive insights, leading to improved health outcomes and a better quality of life.

1. This research enables the early prediction of BG fluctuations, allowing for timely interventions to prevent critical conditions like DKA and severe hypoglycemia. It enhances patient safety and supports real-time self-management.
2. By providing predictive information, the research empowers healthcare professionals with models to formulate treatment plans more accurately. This enhances individualization of care and enhances diabetic treatment.
3. Population-level implementation of predictive models can make diabetes control better overall, thus reducing complication rates and enhancing public health outcomes.
4. Prevention of emergency events and hospitalization by improving predictability and controllability conserves the economic cost of healthcare systems, making delivery of healthcare feasible.
5. The models are also potentially reusable as learning aids, raising patient and population awareness and cognitive understanding of glucose dynamics, allowing proactive and informed self-management behaviors.
6. The predictive data so created can be used as a foundation for ongoing research, spurring innovations in diabetes management, model refinements, and drug development.

In total, this research makes a multi-purposed contribution by enhancing the health status of patients, aiding clinical decision-making, contributing to public health, reducing the cost of care, aiding consumers, and assisting additional scientific research in caring for diabetes.

PROPOSED METHODOLOGY AND RESULT

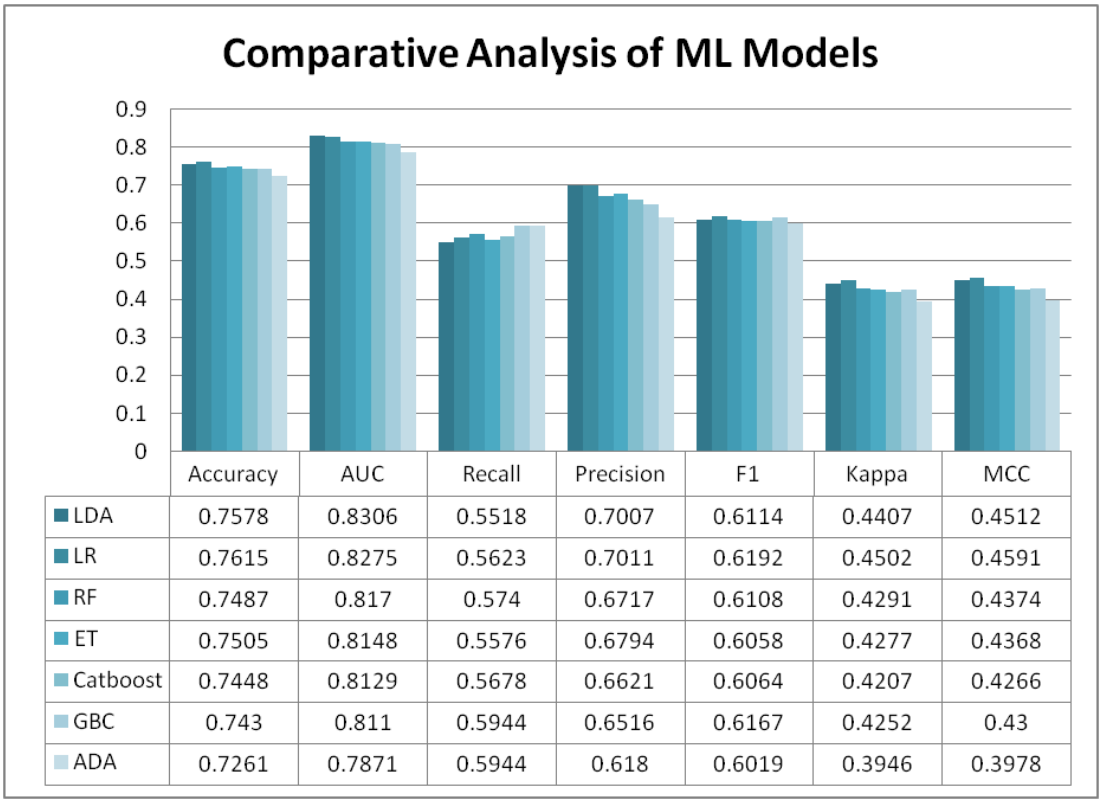
This research successfully developed and validated a dual-pronged strategy for diabetes management, which includes a highly accurate diabetes prediction model and a robust blood glucose level (BGL) forecasting system.

5.1. Ensemble Model for Diabetes Prediction Dataset

The research offers an exhaustive approach to predicting diabetes using the Pima Indians Diabetes Dataset, donated by the National Institute of Diabetes and Digestive and Kidney NIDDK Diseases. This dataset includes diagnosis data on females aged 21 years and above, eight input variables like BMI, BG, insulin, blood pressure, skinfold thickness, pregnancies, age, and diabetes pedigree function. The outcome variable indicates the presence or absence of diabetes. Initial data analysis revealed a high proportion of zero values in the *SkinThickness* and *Insulin* attributes, which were addressed through preprocessing. Linear and standard scaling methods were initially applied, but due to skewed distributions, QuantileTransformer was used for nonlinear scaling to normalize the data.

Feature selection based on correlation showed *Glucose*, *BMI*, *Pregnancies*, and *Age* to be strongly associated with diabetes prediction. The PyCaret framework was employed to build and compare multiple classification models, including CatBoost, Random Forest, Logistic Regression, LDA, and Gradient Boosting. CatBoost emerged as the top performer as shown in figure 1.

Feature importance analysis using SHAP values confirmed *Glucose*, *BMI*, and *Age* as critical predictors, while *SkinThickness* and *Blood Pressure* contributed the least.



	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
LDA	0.7578	0.8306	0.5518	0.7007	0.6114	0.4407	0.4512
LR	0.7615	0.8275	0.5623	0.7011	0.6192	0.4502	0.4591
RF	0.7487	0.817	0.574	0.6717	0.6108	0.4291	0.4374
ET	0.7505	0.8148	0.5576	0.6794	0.6058	0.4277	0.4368
Catboost	0.7448	0.8129	0.5678	0.6621	0.6064	0.4207	0.4266
GBC	0.743	0.811	0.5944	0.6516	0.6167	0.4252	0.43
ADA	0.7261	0.7871	0.5944	0.618	0.6019	0.3946	0.3978

Fig 1: Comparative analysis of ML models

To improve prediction accuracy, ensemble learning techniques—Stacking, Soft Voting, and Hard Voting—were applied. Among them, Soft Voting provided the most balanced and highest accuracy of 92.71%, followed by Hard Voting at 91.15% and Stacking at 89.58%. The final model was further optimized using hyperparameter tuning and calibrated using logistic regression. It was then finalized and deployed to support both individual and batch predictions. The comprehensive framework integrates all stages—preprocessing, model comparison, interpretability, ensemble optimization, and deployment—into a unified system as shown in figure 2.

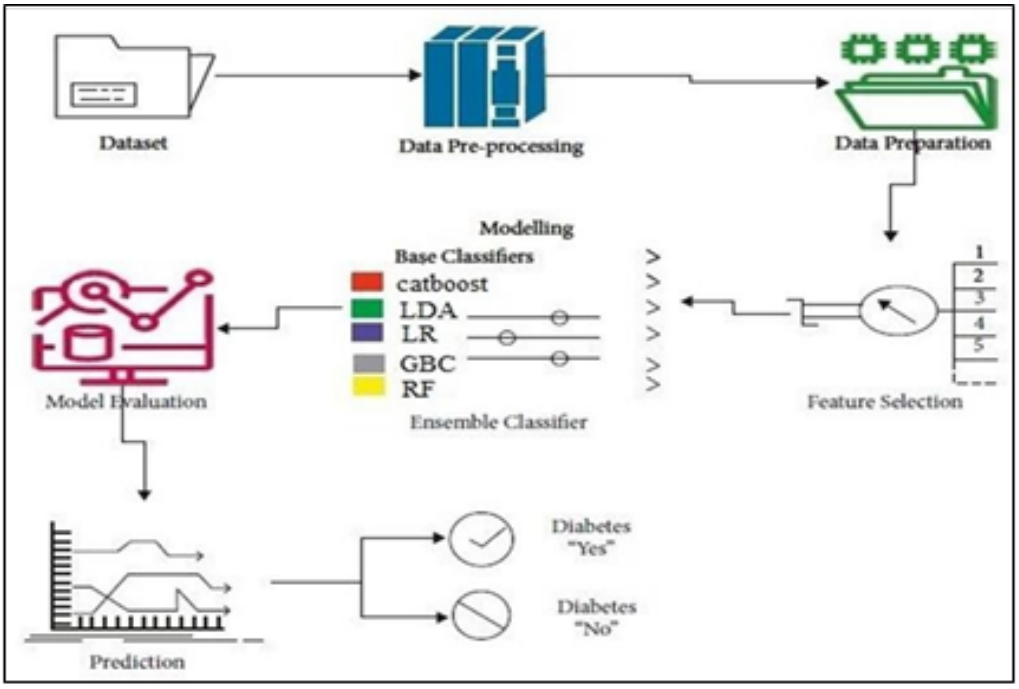
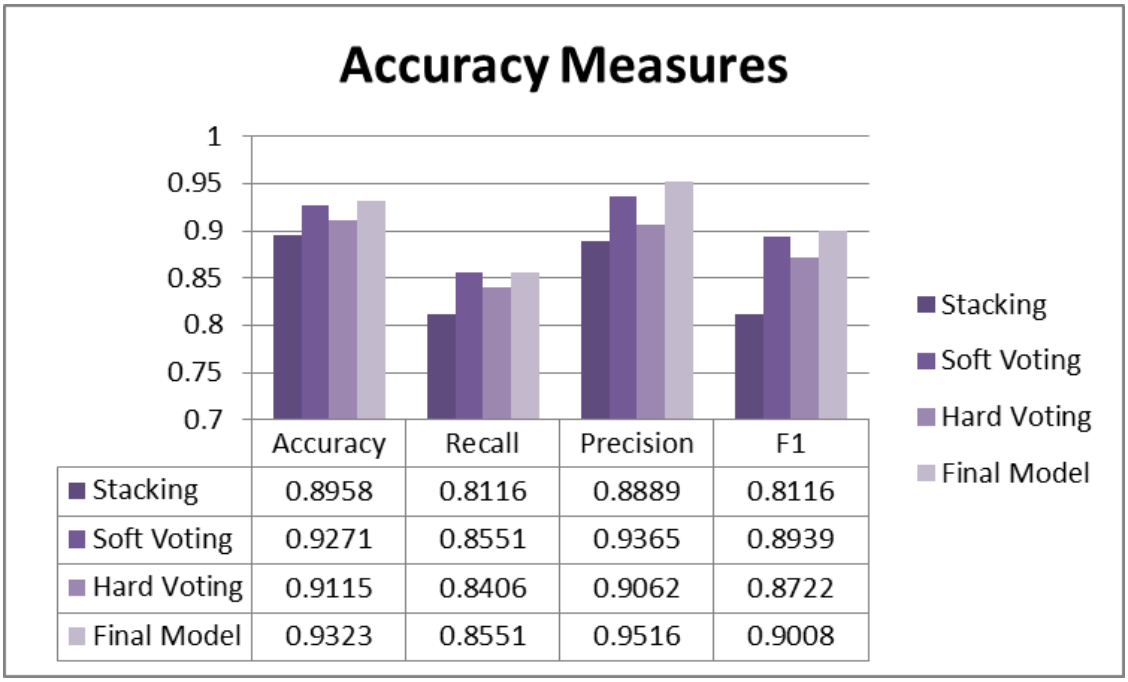


Fig 2: Proposed System Architecture

The final ensemble model, based on Soft Voting, achieved 93.23% accuracy, 85.51% recall, 95.16% precision, 90.08% F1 score, and an AUC of 92.56% as shown in figure 3, outperforming existing ensemble methods such as the one proposed by Saloni Kumari et al [12]. This confirms the proposed method as a reliable, interpretable, and high-performing solution for diabetes prediction.



	Accuracy	Recall	Precision	F1
Stacking	0.8958	0.8116	0.8889	0.8116
Soft Voting	0.9271	0.8551	0.9365	0.8939
Hard Voting	0.9115	0.8406	0.9062	0.8722
Final Model	0.9323	0.8551	0.9516	0.9008

Fig 3: Performance of Ensemble Classifiers on Pima Dataset

5.2 Blood Glucose Level Prediction

The predictive model of BG levels utilized ensemble ML and DL models. For individual patient data, the augmented dataset was utilized for testing and model development.

This paper describes a complete methodology for diabetic patient data processing, prediction of blood glucose, and forecast of complications. The methodology combines data acquisition, preprocessing, augmentation, and application of various ML and DL architectures. Our method used a data-driven model for prediction of BG and patient-specific clustering as follows:

The work uses the D1NAMO dataset, a multimodal physiological dataset of 9 Type 1 Diabetes Mellitus (T1DM) patients' data. The data were recorded using the Zephyr BioHarness 3 wearable device for physiological signals such as ECG, breathing, and 3D acceleration, and glucose levels from the iPro2 CGM sensor in 5-minute resolution. Heart rate, respiratory rate, posture, and insulin doses were time-stamped sensor reading synchronized and resampled in 1-second frequency for preprocessing.

Redundant timestamps were removed, missing values imputed, and irrelevant features discarded. To address limited data, a data augmentation technique was employed by replicating patient data, shifting dates, and adding Gaussian noise (5% of standard deviation), resulting in simulated datasets representing 20 patients over 20 days each.

BGL prediction utilized both ensemble and DL models. Ensemble methods included Random Forest Regressor, Gradient Boosting Regressor, and Ridge Regression, combined via Stacking and Voting Regressors. Input features included physiological signals, activity levels, and insulin dosages, with glucose as the target. Data was scaled, and an 80/20 train-test split was applied. For DL, three custom architectures were developed: a multi-layer LSTM, CNN-LSTM hybrid, and GRU-LSTM hybrid. These models integrated temporal and spatial feature learning, with dropout layers added for regularization. All models were trained for 10 epochs with a batch size of 32 using the Adam optimizer and MSE loss. Forecasts were evaluated at 30, 60, and 90-minute horizons using RMSE and MARD.

To personalize prediction, patient-specific glucose clustering classified readings into hypoglycemia (<4 mmol/L), euglycemia (4–7 mmol/L), and hyperglycemia (>7 mmol/L), aligning with clinical guidelines. Additionally, diabetic complication prediction was performed using features like glucose, heart rate, and core temperature. A custom function labeled each record into complication stages (e.g., normal, prediabetes, hyperglycemia levels). Random Forest and SVM classifiers were employed, and performance was evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrices. Feature importance was visualized for interpretability. Overall, this integrated framework demonstrates a robust, scalable, and clinically relevant system for BG forecasting and complication management. Figure 4 outlines a systematic pipeline for building and evaluating a BG prediction system as described above.

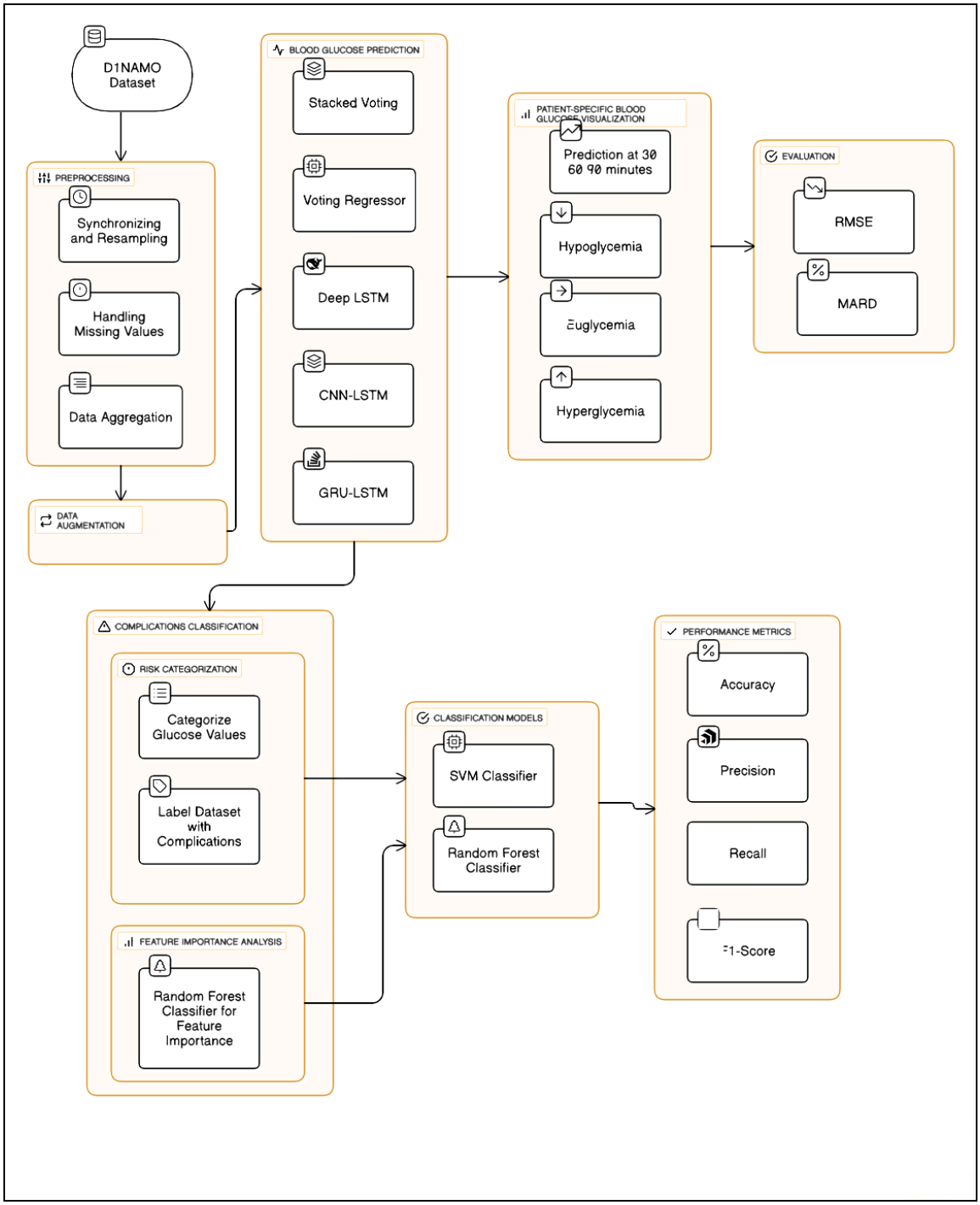
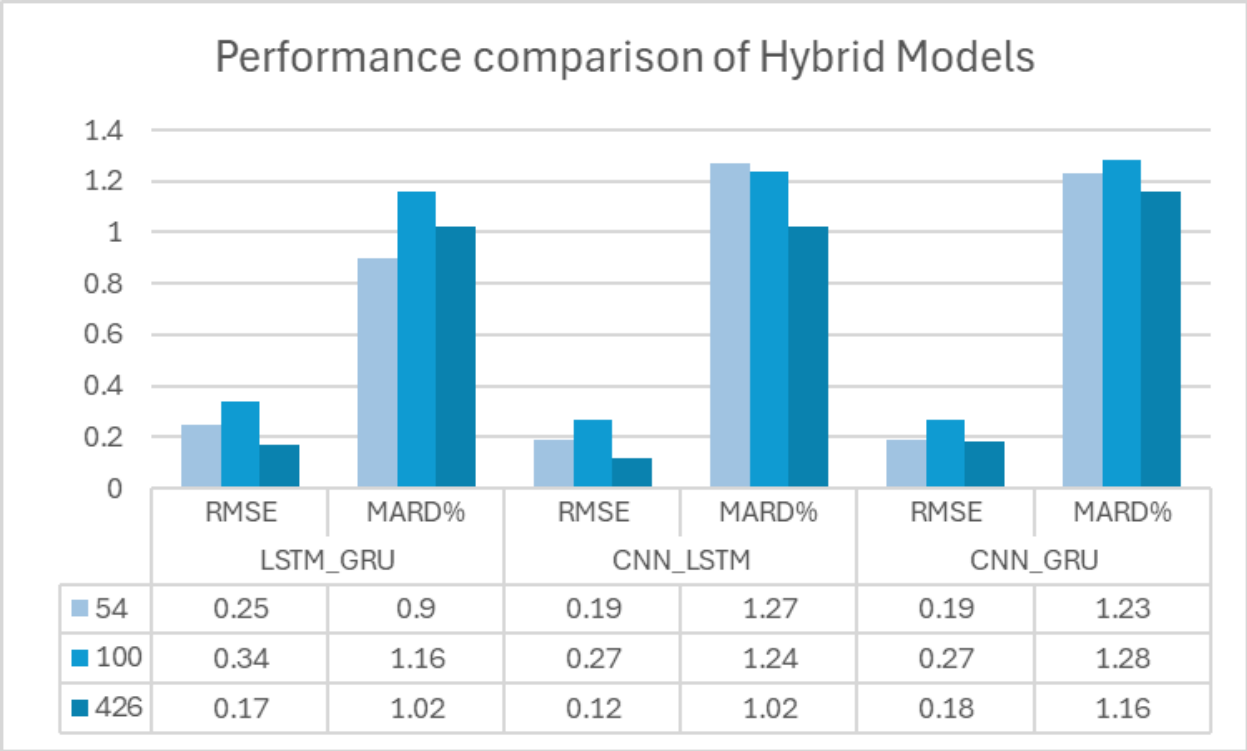


Fig 4: Blood Glucose Prediction Architecture

After training and prediction, each model's results were printed in terms of RMSE and MARD for 30, 60, and 90-minute intervals. This enables a comparative assessment of model robustness and predictive accuracy over varying forecast horizons. The outputs

help identify which architecture offers the best trade-off between short-term responsiveness and long-term reliability in glucose prediction tasks.

Figure 5 represents the performance metrics of all BG prediction models. This table provides a direct comparison of the accuracy of Stacked Model, Voting Model, DLSTM LSTM -GRU and CNN-LSTM in predicting BGL for 20 different patients. For each model and each patient, Root Mean Square Error (RMSE) and Mean Absolute Relative Difference (MARD) critical error metrics are reported:



	LSTM_GRU		CNN_LSTM		CNN_GRU	
	RMSE	MARD%	RMSE	MARD%	RMSE	MARD%
54	0.25	0.9	0.19	1.27	0.19	1.23
100	0.34	1.16	0.27	1.24	0.27	1.28
426	0.17	1.02	0.12	1.02	0.18	1.16

Fig 5: Hybrid Models comparison for prediction of blood glucose prediction

Stacked Model, particularly with lower MARD values for several patients. The DLSTMmodel demonstrates remarkably low RMSE and MARD values across almost all patients.DLSTM maintains a high level of accuracy across different patients, indicating its robustness.

The following graph (figure 6) represents a time-series plot of the true BG values for Patient 15 with the predictions made by the CNN-LSTM Model at different future horizons (30, 60, and 90 samples/minutes). Similarly the graphs with future horizons and analysis was conducted for all remaining 20 patients for all models.

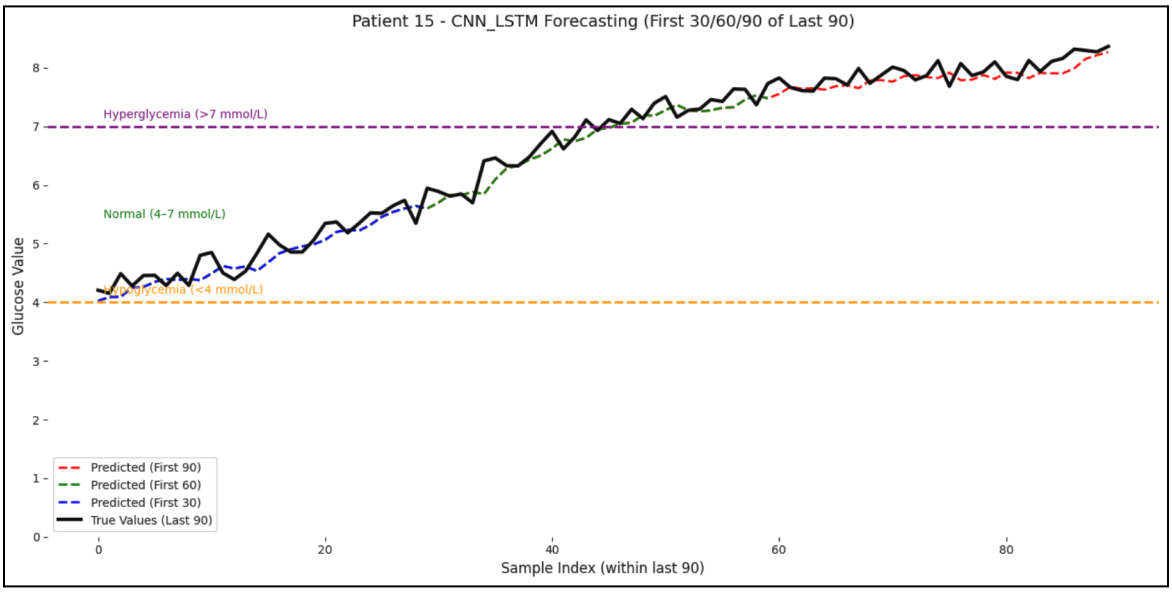


Fig 6: Blood Glucose Prediction for Subject 15 (pH=30,60,90 Mins)

All three models excel at avoiding false hypoglycemia predictions. The DLSTM model shows a much more precise and less predictive tracking of the true glucose values when they are elevated (>7 mmol/L). The threshold lines helped to assess the model can accurately predict when the patient enters hypoglycemic or hyperglycemic states, which is crucial for timely interventions. In this particular graph, the true values for Patient 2 primarily stay between the 4 and 7 mmol/L thresholds, with some periods approaching or briefly crossing the 7 mmol/L mark. The very low RMSE and MARE values, making the DLSTM model the most reliable choice for predicting glucose trends and specific levels for Patient ID 2, which is crucial for informed clinical decisions. This research significantly advances personalized diabetes management by introducing a highly accurate and patient-specific glucose forecasting system.

CONCLUSION

This work was able to design and validate a two-pronged strategy for diabetes management: an extremely accurate model for diabetes prediction and an efficient blood glucose level (BGL) forecasting system. With the use of cutting-edge machine learning and deep learning methods, this work describes a systems-level, data-driven approach that has the potential to shift diabetes management from reactive to proactive and personalized.

Our initial contribution, the diabetes prediction model, was learnt on the Pima Indians Diabetes Dataset. Through extensive experimentation with different classifiers and ensemble techniques, we determined that a Soft Voting ensemble model could predict diabetes at a rate of 93.23%. The model, capitalizing on the unique strengths of the base classifiers, outperformed current methods, validating that an effectively designed ensemble method is capable of dramatically boosting the accuracy of diabetes diagnosis.

The second part of the work involved short-term prediction of BGL in Type 1 diabetes patients. Using the D1NAMO dataset, we tried training and testing some models, both deep learning and ensembling. The Deep Long Short-Term Memory (DLSTM) model performed extremely well and outperformed all other models at all times, with very low values of Root Mean Square Error (RMSE) and Mean Absolute Relative Difference (MARD) in all patients and prediction horizons (30, 60, and 90 minutes). Its precision and reasonableness are of utmost importance in avoiding harmful hyperglycemic and hypoglycemic events. In addition, our patient-specific clustering and complication prediction models introduce an important level of personalized care, allowing for customized interventions and risk estimation.

The integrated model proposed here is a significant step forward in patient-specific diabetes management. By the accurate prediction of the long-term diabetes risk and short-term blood glucose variability, our system offers a useful tool for the evidence-based choice for clinicians and patients. It enhances health outcomes, reduces healthcare costs by preventing emergency events, and forms the basis for further innovations in diabetes management.

ORGANIZATION OF THESIS

TABLE OF CONTENTS

Sr. No.	Topic	Page No.
	Abstract……..	vi
	List of Figures ……..	xiii
	List of Tables …….	xv
	Abbreviations…….	xvi
CHAPTER 1: INTRODUCTION		1-13
1.1	Overview…….	1
1.2	Background of Diabetes Mellitus……	3
	1.2.1 Types and Causes of Diabetes……	3
	1.2.2 Risk Factors……	4
	1.2.3 Diabetes Complications…….	4
1.3	Research Motivation …….	6
1.4	Problem Statement …….	7
1.5	Scope of the Research…….	8
1.6	Significance of the study…….	10

1.7 Research Contributions ……. 10

1.8 Thesis Organization ……... 11

CHAPTER 2: LITERATURE REVIEW 14-33

2.1 Review on Diabetes Prediction……. 15

2.2 Review on Blood Glucose Level Prediction…… 21

2.3 Review on Prediction of Diabetes Complications…… 27

2.4 Summary of Findings……... 28

2.4.1 ML and Ensemble-Based Approaches…… 28

2.4.2. Deep Learning Approach for Blood Glucose Prediction Models…….. 29

2.5 Research Gap…… 31

2.6 Research Objectives……. 32

CHAPTER 3: DATASETS 33-41

3.1 Dataset Overview……. 33

3.1.1 Pima Indian Diabetes Dataset: Dataset 1……. 34

3.1.2 Neha Tigga Dataset: Dataset 2……. 36

3.1.3 Diabetes Dataset: Dataset 3…….. 37

3.2 Tools and Techniques used for data collection…….. 38

3.2.1 Diabetes Technology…….. 38

3.2.2 Continuous Glucose Monitor (CGM)……. 39

3.3 Dataset Used for BGL Prediction: D1NAMO Dataset…….. 39

CHAPTER 4: TECHNIQUES FOR DIABETES PREDICTION AND BLOOD GLUCOSE LEVEL PREDICTION 42-56

4.1 Machine Learning Methods for Diabetes Prediction……... 42

4.1.1 Decision Tree…… 42

4.1.2 Naive Bayes ……. 43

4.1.3 Logistic Regression ……. 44

4.1.4 AdaBoost…….. 44

4.1.5 Random Forest ……. 44

4.1.6 K-Nearest Neighbour…… 45

4.1.7 Linear Discriminant Analysis……... 45

4.1.8 Gradient Boosting Classifier……. 45

4.1.9 CatBoost Classifier……... 45

4.2 Ensemble Methods for Diabetes Prediction…… 47

4.2.1 Stacking……. 47

4.2.2 Voting Classifiers…… 47

4.2.2.1 Soft Voting…….. 48

4.2.2.2 Hard Voting…… 49

4.3 Deep Learning and Hybrid Methods for Blood Glucose Prediction…….. 51

4.3.1 Long Short-Term Memory…… 51

4.3.2 Gated Recurrent Unit…… 53

4.3.3. Convolutional Neural Network…… 55

CHAPTER 5: EXPERIMENTAL SETUP 57-96

Model A: Comparative Analysis of Different Diabetes Datasets Using ML Techniques 57

5.1 Data Preprocessing…….. 57

5.2 Algorithmic Framework for Diabetes Prediction…… 59

5.3 Proposed Architecture for Diabetes Prediction…… 60

Model B: Diabetes Prediction Using Ensemble Classifiers 62

5.4 Exploratory Data Analysis…….. 63

5.5 Model Creation and Training ……. 68

5.6	Framework for Diabetes Prediction using Ensemble Models….	69
5.7	Comparative Analysis of ML models…….	72
5.8	Model Interpretations……...	73
	Model C: Prediction of Diabetes Blood Glucose Level	74
5.9	Data Preprocessing……..	75
5.10	Data Augmentation…….	77
5.11	Framework for Blood Glucose Prediction and Diabetes Complication Classification…….	78
5.12	Blood Glucose Prediction Models……	81
5.12.1	Architecture 1—BGL Prediction Without Augmentation……	82
5.12.2	Architecture 2—CGM Experiments: 20 Patients → 20 Synthetic Patients……..	83
5.12.3	Architecture 3—Scaling to Larger CGM-Only Cohorts (20, 54, and 100 Patients)…….	84
5.12.4	Architecture 4— Multimodal Physiological Input Experiments: 50, 100, and 500 Patients ……	84
5.13	Model Architectures and Summary…….	87
	5.13.1 Deep Long Short-Term Memory ……	87

5.13.2 Convolutional Neural Network-Long Short-Term Memory…….. 89

5.13.3 Long Short-Term Memory—Gated Recurrent Unit …… 90

5.13.4 Convolutional Neural Network-Gated Recurrent Unit …… 91

5.14 Training and Evaluation…….. 93

5.15 Patient-Specific Blood Glucose Clustering……. 94

5.16 Feature Importance Analysis…….. 95

5.17 Diabetes Complications Classification…… 96

CHAPTER 6: RESULTS AND DISCUSSION 97-119

Model A: Comparative Analysis of Different Diabetes Datasets Using ML Techniques 97

6.1 Experimental Results ……. 97

6.2 Comparative Analysis……. 99

6.3 Summary……. 100

Model B: Diabetes Prediction Using Ensemble Classifiers 100

6.4 Experimental Results ……. 100

6.4.1 Calibrated Model…….. 103

6.4.2 Finalized Model…… 103

6.5 Comparative Analysis…….. 106

6.6 Summary……. 107

Model C: Prediction Of Diabetes Blood Glucose Level

6.7 Architecture 1 Results……. 108

6.8 Architecture 2 Results……. 109

6.9 Architecture 3 Results……. 110

6.10 Architecture 4 Results…….. 111

6.11 Model Validation……. 112

6.12 Results of Diabetes Classification…… 114

6.13 Comparative Analysis…….. 115

6.14 Summary……. 117

CHAPTER 7: CONCLUSION AND FUTURE SCOPE 120-121

7.1 Conclusion…… 120

7.2 Future Scope…… 121

References 122-131

Appendix A (List of Research Publications) 132

Appendix B (Patents) 133

Appendix C (Plagiarism Reports) 134

REFERENCES

1. Saiful, Islam “A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining” , Healthcare (Basel). 2018 Jun; 6(2): 54, doi: 10.3390/healthcare6020054.
2. A. Hussain and S. Naaz, “Prediction of diabetes mellitus: comparative study of various machine learning models,” in Proceedings of the International Conference on Innovative Computing and Communications, pp. 103–115, Springer, Delhi, India, January 2021.
3. Zhang LM. Genetic deep neural networks using different activation functions for financial data mining. In: Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015; 2015.p. 2849–51. https://doi.org/10.1109/BigData.2015.7364099.
4. Huma Naz, and Sachin Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset", J Diabetes Metab Disord. 2020 Jun; 19(1): 391–403,doi:10.1007/s40200-020-00520-5
5. Swapna G, Soman K.P., Vinayakumar R," Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals", Procedia Comput. Sci. 2018, 132, 1253–1262.
6. Huaping Zhou, Raushan Myrzashovam, Rui Zheng, "Diabetes prediction model based on an enhanced deep neural network", July 2020 EURASIP Journal on Wireless Communications and Networking 2020(1), DOI:10.1186/s13638-020-01765-7.
7. Hrushikesh N. Mhaskar, Sergei V. Pereverzyev,"A Deep Learning Approach to Diabetic Prediction", Front. Appl. Math. Stat., 14 July 2017 Sec. Mathematics of Computation and Data Science, Volume 3 - 2017 | https://doi.org/10.3389/fams.2017.00014.
8. Akihiro Nomura, Masahiro Noguchi, Mitsuhiro Kometani, Kenji Furukawa, Takashi Yoneda," Artificial Intelligence in Current Diabetes Management and Prediction", Curr Diab Rep, 2021 Dec 13;21(12):61. DOI: 10.1007/s11892-021-01423-2.
9. E.I Georga, V. Protopappas, A. Guillen, G. Fico, D. Ardigo, M. Arredondo, T. P. Exarchos, D. Polyzos, D.I. Fotiadis, “ Data Mining for Blood Glucose Prediction and Knowledge Discovery in Diabetic Patients: The METABO Diabetes Modeling and Management System,” in 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, pp. 5633—5636, 2009.

10. Radhi Meenakshi S, " Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural networks" In: 2016 International Conference on Computing for Sustainable Global Development (INDIACom); 2016;3107–11.
11. Neha Sharma, Ashima Singh, "Diabetes Detection and Prediction Using Machine Learning/IoT: A Survey", ICAICR 2018, Shimla, India, July 14–15, 2018, DOI:10.1007/978-981-13-3140-4_42D.
12. S. Malik, S. Harous, and H. E. Sayed, “Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women,” in Proceedings of the International Symposium on Modelling and Implementation of Complex Systems, pp. 95–106, Springer, Batna, Algeria, October 2020.
13. Pradhan, Nitesh & Rani, Geeta & Dhaka, Vijaypal & Poonia, Ramesh. (2020). Diabetes prediction using artificial neural networks. DOI:10.1016/B978-0-12-819061-6.00014-8.
14. Neha Prerna Tiggaa, Shruti Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods”, International Conference on Computational Intelligence and Data Science (ICCIDS 2019), ScienceDirect, Procedia Computer Science 167 (2020) 706–716.
15. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", Informatics in Medicine Unlocked, Volume 10, 2018, Pages 100-107.
16. Jobeda Jamal Khanam, Simon Y. Foo,"A comparison of machine learning algorithms for diabetes prediction", February 2021ICT Express 7(1), DOI:10.1016/j.icte.2021.02.004.
17. Ismail, L., Materwala, H., Tayefi, M. *et al.* Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. *Arch Computat Methods Eng* 29, 313–333 (2022). https://doi.org/10.1007/s11831-021-09582-x.
18. Ali Mohebbi, Tinna B. Aradottir , Alexander R. Johansen , Henrik Bengtsson , Marco Fraccaro , Morten Mørup, “A Deep Learning Approach to Adherence Detection for Type 2 Diabetics”, 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) , 978-1-5090-2809-2/17, IEEE.
19. Gómez-Castillo, N.Y. *et al.* (2022). A Machine Learning Approach for Blood Glucose Level Prediction Using a LSTM Network. In: Narváez, F.R., Proaño, J., Morillo, P.,

Vallejo, D., González Montoya, D., Díaz, G.M. (eds) Smart Technologies, Systems and Applications. SmartTech-IC 2021. Communications in Computer and Information Science, vol 1532. Springer, Cham. https://doi.org/10.1007/978-3-030-99170-8_8.

20. Bhimireddy, Ananth, Sinha, Priyanshu, Oluwalade, Bolu, Gichoya, Judy W, & Purkayastha, Saptarshi. *Blood Glucose Level Prediction as Time-Series Modeling using Sequence-to-Sequence Neural Networks*. *CEUR workshop proceedings*, (). Retrieved from https://par.nsf.gov/biblio/10188463.
21. Munoz-Organero M. Deep Physiological Model for Blood Glucose Prediction in T1DMPatients. Sensors (Basel). 2020 Jul 13;20(14):3896. doi: 10.3390/s20143896. PMID:32668724; PMCID: PMC7412558.
22. Zaidi, S.M.A., Chandola, V., Ibrahim, M. *et al.* Multi-step ahead predictive model for blood glucose concentrations of type-1 diabetic patients. *Sci Rep* 11, 24332 (2021). https://doi.org/10.1038/s41598-021-03341-5).
23. D. Y. Kim, D. S. Choi, A. R. Kang, J. Woo, Y. Han, and S. W. Chun, "Intelligent ensemble deep learning system for blood glucose prediction using genetic algorithms," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/7902418.
24. Jaouher Ben Ali, Takoua Hamdi, Nader Fnaiech, Véronique Di Costanzo, Farhat Fnaiech, Jean-Marc Ginoux, (Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network, Biocybernetics and Biomedical Engineering,Volume 38, Issue 4, 2018, Pages 828-840, ISSN 0208-5216, https://doi.org/10.1016/j.bbe.2018.06.005.
25. Jian Y, Pasquier M, Sagahyroon A, Aloul F. A Machine Learning Approach to Predicting Diabetes Complications. Healthcare (Basel). 2021 Dec 9;9(12):1712. doi: 10.3390/healthcare9121712. PMID: 34946438; PMCID: PMC8702133.
26. Ravaut, M., Sadeghi, H., Leung, K.K. *et al.* Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *npj Digit. Med.* 4, 24 (2021). https://doi.org/10.1038/s41746-021-00394-8.
27. M. V. Kabyshev and S. V. Kovalchuk, "Development of Personalized Mobile Assistant for Chronic Disease Patients: Diabetes Mellitus Case Study", *Procedia Computer Science*, vol. 156, pp. 123-33, 2019

28. Seo W, Park SW, Kim N, Jin SM, Park SM. A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study. Comput Methods Programs Biomed. 2021 Nov;211:106424. doi: 10.1016/j.cmpb.2021.106424. Epub 2021 Sep 20. PMID: 34598081.
29. Ganjar Alfian , Muhammad Syafrudin Muhammad Fazal Ijaz , M. Alex Syaekhoni , Norma Latif Fitriyani and Jongtae Rhee " A Personalized Healthcare Monitoring System for Diabetic Patients by Utilizing BLE-Based Sensors and Real-Time Data Processing", Sensors 2018, 18, 2183; doi:10.3390/s18072183, MDPI.
30. Robert A. Sowah , Adelaide A. Bampoe-Addo, Stephen K. Armoo, Firibu K. Saalia, Francis Gatsi, and Baffour Sarkodie-Mensah, "Design and Development of Diabetes Management System Using Machine Learning" Hindawi International Journal of Telemedicine and Applications Volume 2020, Article ID 8870141
31. Ümit Şentürk, Gökhan Adıgüzel, Kemal Polat, Hybrid Mamba-MoE model for non-invasive blood glucose prediction, Computers and Electrical Engineering, Volume 127,PartA,2025,110549,ISSN0045-7906, https://doi.org/10.1016/j.compeleceng.2025.110549.

LIST OF RESEARCH PUBLICATIONS

1. Madhuri Kawarkhe, Parminder Kaur, Prediction of Diabetes Using Diverse Ensemble Learning Classifiers, Procedia Computer Science, Volume 235, 2024, Pages 403-413, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2024.04.040.
2. Kawarkhe Madhuri, and Parminder Kaur. "Comparative analysis of machine learning algorithms on different diabetes datasets." In *American Institute of Physics Conference Series*, vol. 3188, no. 1, p. 080026. 2024, https://doi.org/10.1063/5.0240228.
3. Madhuri Kawarkhe and Dr. Parminder Kaur, “Prediction of Diabetes Symptoms and Blood Glucose Levels using Wearable Devices”, The first international conference on NEXT-GEN AI and ML 2025.
4. Kawarkhe Madhuri & Kaur Parminder. (2025). Prediction of Blood Glucose Levels and Diabetes Complications Using Wearables for Type1 Diabetics. Journal of Neonatal Surgery. 14. 900-908. 10.63682/jns.v14i20S.5368, https://doi.org/10.63682/jns.v14i20S.5368.
5. Madhuri Kawarkhe & Dr. Parminder Kaur, “Scope of Research in Diabetes and a case study using Lazy Prediction on the pima dataset”, Frontiers of Science and Technology: Integrating Innovations for Sustainable Future Vol - 1, ISBN-978-93-94174-76-4, https://doi.org/10.20546/EP-978-93-94174-76-4.

Patent Published

1. A device to predict Blood Glucose Levels and Diabetes Complications

Office of the Controller General of Patents, Designs & Trade Marks
Department for Promotion of Industry and Internal Trade
Ministry of Commerce & Industry,
Government of India

INTELLECTUAL PROPERTY INDIA
PATENTS | DESIGNS | TRADE MARKS | GEOGRAPHICAL INDICATIONS

Application Details	
APPLICATION NUMBER	202521015527
APPLICATION TYPE	ORDINARY APPLICATION
DATE OF FILING	22/02/2025
APPLICANT NAME	**1 . Madhuri B. Kawarkhe** **2 . Dr. Parminder Kaur**
TITLE OF INVENTION	A Device to Predict Blood glucose levels and Diabetes complications
FIELD OF INVENTION	CHEMICAL
E-MAIL (As Per Record)	rpyadav@sr4ipr.in
ADDITIONAL-EMAIL (As Per Record)	sr4iprpartners@gmail.com
E-MAIL (UPDATED Online)	
PRIORITY DATE	
REQUEST FOR EXAMINATION DATE	--
PUBLICATION DATE (U/S 11A)	07/03/2025